Cloud Computing and Data Science
http://ojs.wiserpub.com/index.php/CCDS/

Research Article

Machine Learning Analysis of Factors Contributing to Diabetes Development

Edgar Ceh-Varela*, Larry Maes, Sarbagya Ratna Shakya

Department of Mathematical Sciences, Eastern New Mexico University, Portales, NM, USA
Email: Eduardo.Ceh@enmu.edu

Received: 8 October 2023; **Revised:** 18 December 2023; **Accepted:** 19 December 2023

Abstract: Diabetes is a chronic condition that affects how the body processes blood sugar. Early diagnosis and management of diabetes are essential for preventing these complications. Machine Learning (ML) techniques offer an effective means to accurately diagnose diabetes by identifying key risk factors and developing predictive models. In this study, we assess the performance of 11 ML algorithms on four diabetes prediction datasets, considering the top 2, top 3, and all attributes. Through k-fold cross-validation, we ensure robust and generalizable results. We use a set of standard evaluation metrics such as accuracy, precision, recall, f1-score, and Receiver Operating Characteristic curve (ROC_AUC). Our analysis aims to determine the optimal number of features and assess how performance changes with feature additions. Notably, some ML classifiers achieve satisfactory classification and predictive abilities using only the top 2 or 3 features. Furthermore, varying dataset performances across algorithms highlight the need for assessing multiple models to identify the most suitable one. These findings enable the creation of dependable models that enhance patient outcomes by leveraging effective algorithms and pertinent features.

***Keywords*:** diabetes, machine learning, predictive models, performance evaluation, feature selection

1. Introduction

Diabetes is a chronic disease affecting millions worldwide, with an increasing prevalence over the past few decades [1-2]. It is a complex metabolic disorder that occurs when the body cannot regulate blood glucose levels due to either a lack of insulin production or an inability to use insulin effectively. If left unmanaged, diabetes can lead to various complications, including cardiovascular disease, kidney failure, blindness, and nerve damage [3].

As such, predicting and preventing diabetes is a significant public health concern. Machine learning (ML) techniques have shown considerable promise in this area, as they allow for the development of accurate predictive models based on complex relationships between various data points. These models are useful in identifying patterns in extensive patient data that humans would probably miss. Recently, ML techniques have been applied to diabetes prediction, resulting in numerous studies exploring different algorithms and datasets [4-8].

The most commonly used ML techniques in diabetes prediction include Logistic Regression, Decision Trees, Support Vector Machines (SVM), Random Forest, and Artificial Neural Networks (ANN), all of which have shown varying degrees of success.

Various factors, including the size and quality of the dataset, the choice of ML algorithm, and the selection of

DOI: https://doi.org/10.37256/ccds.5120243751

relevant features, influence the accuracy of these ML models. Combining different dataset features as predictors can further improve the performance of these techniques. However, identifying the most important features for predicting diabetes remains a challenging problem, and different datasets may require different features to achieve the highest accuracy.

Hence, this study is motivated to assess the efficacy of diverse ML algorithms in predicting diabetes, employing four distinct datasets. Our focus includes evaluating various ML algorithms for diabetes prediction based on the top 2, top 3, and all dataset features. Additionally, we aim to analyze how different performance metrics, including accuracy, precision, recall, f1-Score, and ROC_AUC influence model evaluation.

The four datasets we will use in this study are the Mendeley dataset, the Pima Indians Diabetes (PID) dataset, the Diabetes Early Stage (DES) dataset, and the Vanderbilt dataset (see Section 4.1). Each dataset has unique characteristics, including the number of instances, features, target variables, and differences in the populations they represent. By exploring the performance of various ML algorithms on different datasets, we aim to provide a comprehensive evaluation of the effectiveness of these techniques for predicting and managing diabetes.

This study makes the following important contributions:

a) It rigorously benchmarks predictive performance for diabetes risk across 11 diverse ML models and four real-world datasets, facilitating an accurate assessment.

b) It provides interpretable feature importance rankings, unveiling actionable risk factors.

c) It offers a more comprehensive and multi-faceted examination.

d) Its rigorous methodology corroborates the reliability and robustness of our findings, surpassing studies primarily concentrating on a few algorithms and requiring more extensive feature scrutiny.

The results of our study hold significant implications for advancing the application of precise and effective diabetes prediction models. These advancements can improve patient outcomes and mitigate complications associated with the disease. Moreover, our research contributes to the expanding knowledge of utilizing ML techniques for predicting and managing chronic diseases. This area of research has gained prominence in recent years, highlighting the relevance and impact of our study within this evolving field.

The remainder of this paper is organized as follows: Section 2 presents the literature related to our research. Section 3 details the proposed method. The results of applying the ML techniques to the different datasets are presented in Section 4. Finally, our conclusions are presented in Section 5.

2. Literature review

Diabetes is a chronic condition affecting many people globally, and its prediction and prevention are crucial for public health. This disease is characterized by high blood sugar levels, also known as hyperglycemia, and can cause various complications if not managed properly. Age, ethnicity, family history, low socioeconomic level, obesity, metabolic syndrome, cardiac complications, food intake, and some bad lifestyle choices are the main risk factors for diabetes [9-10]. The World Health Organization predicts that by 2040, the number of individuals living with diabetes will reach 642 million, which translates to one out of every ten adults [1]. This alarming statistic emphasizes the need for effective approaches to tackling the increasing incidence of diabetes.

One promising method for diabetes prediction is the use of machine learning methods [4]. Machine learning (ML) is a branch of artificial intelligence (AI) that enables computer systems to learn and improve from data without human intervention [11]. The accuracy of diabetes diagnosis and prediction has improved thanks to ML approaches, as demonstrated by encouraging results [12-14]. These methods can analyze large data sets and find patterns humans would miss. For example, ML algorithms can analyze data from electronic health records to predict a person's risk of acquiring diabetes. These algorithms can also be trained on past data, including medical history and lifestyle factors, which enables them to become more accurate predictors as they accumulate more knowledge [8, 15].

In this section, we will explore recent research on ML techniques for diabetes prediction and compare the performance of various ML models using different dataset features.

Alkaragole and Kurnaz [5] studied the precision of different ML methods, including Decision Trees, Naive Bayes, SVM, and hybrid algorithms. They found that combining SVM and Decision Trees was more accurate than the other

algorithms, with a precision of 94% and a sensitivity of 91%. Sneha and Gangil [6] focused on selecting optimal features from the dataset to improve classification accuracy. The researchers used predictive analysis to focus on finding significant features to help with the early detection of Diabetes Mellitus. The results showed that Decision Trees and Random Forest algorithms had the highest specificity of 98.2% and 98%, respectively. Of all the algorithms used, the Naive Bayes algorithm had the highest accuracy of 82.3%. Nuankaew et al. [7] propose a novel method for type 2 diabetes prediction with factors representing personal health conditions. The proposed Average Weighted Objective Distance (AWOD) method is a modification of Weighted Objective Distance (WOD) [16] by applying information gain to reveal significant and insignificant individual factors having different priorities, which are represented by different weights. To validate the proposed method, two open source datasets, Pima Indians Diabetes (Dataset 1) and Mendeley Data for Diabetes (Dataset 2), each containing 392 records, were studied. The comparison results showed that the proposed method provided 93.22% and 98.95% accuracy for Dataset 1 and 2, respectively, which are higher than those provided by other ML methods such as K-Nearest Neighbors (K-NN), SVM, Random Forest, and Deep Learning.

Kaur and Kumari [17] developed five models to detect diabetes using Linear Kernel Support Vector Machine (SVM-Linear), Radial Basis Kernel Support Vector Machine (SVM-RBF), K-NN, ANN, and Multi Dimensional Reduction (MDR) algorithms. Feature selection of the dataset was made using the Boruta wrapper algorithm, which provided an unbiased selection of important features. All the models achieved good results, with SVM-Linear providing the highest accuracy of 89% and precision of 88% for predicting diabetes. K-NN provided the best recall and f1-Score of 90% and 88%, respectively. The AUC values for SVM-Linear and K-NN were 90% and 92%, respectively, indicating that both SVM-Linear and K-NN are optimal classifiers for the diabetic dataset. The research suggested that the Boruta wrapper algorithm can be used for feature selection and is better than choosing attributes manually with less medical domain knowledge.

Similarly, Nguyen et al. [18] propose a wide and deep learning model to predict the onset of type 2 diabetes mellitus (T2DM) using public hospital record data provided by the Practice Fusion EHRs for the United States population. The dataset consists of de-identified electronic health records for 9,948 patients, of which 1,904 have been diagnosed with T2DM. The model was trained using a logistic loss function and a stochastic gradient descent. The imbalance class of the model was handled by the Synthetic Minority Oversampling Technique (SMOTE) for each cross-validation training fold. The results show that the proposed model obtained an accuracy of 84.28% and an area under the Receiver Operating Characteristic curve (ROC_AUC) of 84.13%. Using SMOTE did not improve ROC_AUC but increased sensitivity with a moderate decrease in specificity.

In Sisodia et al. [19] the researchers compared three ML classification algorithms (Decision Trees, SVM, and Naive Bayes) to detect diabetes at an early stage, using the Pima Indians Diabetes (PID) dataset. The algorithms were evaluated on various measures such as precision, accuracy, f1-Score, and recall. Results showed that Naive Bayes had the highest accuracy of 76.3%, verified using Receiver Operating Characteristic (ROC) curves.

Tigga and Garg [20] compared six ML classification methods using a dataset collected through online and offline questionnaires consisting of 18 questions relevant to diabetes and the PID dataset. The results showed that the Random Forest algorithm had the highest accuracy of 94.1% and was the best algorithm in both datasets. They found that the parameters with the highest significance for predicting diabetes were “Age”, “Family diabetes”, “Physically active”, “Regular Medicine”, and “Pdiabetes”. Wu et al. [21] propose a novel model based on data mining techniques for predicting type 2 diabetes mellitus (T2DM). The model comprised the improved K-means algorithm and the Logistic Regression algorithm and was tested on the PID dataset and the Waikato Environment for Knowledge Analysis toolkit. The results showed that the model attained a 3.04% higher prediction accuracy than other studies and was applied to two other diabetes datasets with good performance. This suggests that the model is effective for the realistic health management of diabetes.

Rajendra and Latifi [22] designed a prediction model to detect diabetes using Logistic Regression and explored various techniques to boost performance and accuracy. Two datasets were used, the PID dataset (Dataset 1) and the Vanderbilt dataset (Dataset 2). Feature selection was made using two different methods, and ensemble methods were used to improve performance. The highest accuracy obtained was around 78% for Dataset 1 with Max Voting Ensemble and around 93% for Dataset 2 with Max Voting and Stacking Ensemble. The study concluded that Logistic Regression is an efficient algorithm for building prediction models. The researchers found that data preprocessing, removal of redundant/null values, normalization, cross-validation, feature selection, and ensemble techniques can also improve

accuracy and runtimes.

Similarly, Gill and Pathwat [23] analyzed diabetes symptoms to gather meaningful insights to help health experts make early diagnosis. The researchers used feature selection techniques such as Analysis of variance (ANOVA), mutual information, and genetic algorithm to increase accuracy and reduce overhead and training time. They used Logistic Regression, Naive Bayes, Stochastic Gradient Descent (SGD) Classifier, K-NN, Random Forest, Decision Trees, and SVM algorithms to predict diabetes. Random Forest showed the best accuracy of 93.95%, with Genetic Algorithm as a feature selection technique, selecting "Cholesterol", "Glucose", "Chol/HDL", "Systolic BP", "Weight", and "Hip ratio" as the most important features.

Zout et al. [1] used a Decision Tree, Random Forest, and Neural Network to predict diabetes mellitus using hospital physical examination data in Luzhou, China. They implemented five-fold cross-validation and independent test experiments to verify the models' universal applicability. Similarly, the researchers used Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (MRMR) to reduce the dimensionality. The results showed that Random Forest achieved the highest accuracy of 80.84% when all the attributes were used.

While the presented works demonstrate notable efforts in utilizing ML techniques for diabetes prediction, several weaknesses are evident across these studies. Firstly, there is a lack of consistency regarding dataset characteristics, which hinders the comparability of results. Varied dataset sizes, sources, and features make it challenging to draw conclusive insights or generalize findings. Secondly, the evaluation metrics employed in these studies vary, with some focusing on accuracy, specificity, sensitivity, and area under the curve (AUC), making direct comparisons cumbersome. Additionally, the absence of standardized metrics for assessing model performance across the studies introduces ambiguity and limits the robustness of the comparative analysis. Furthermore, most studies concentrate on a specific subset of ML algorithms, lacking a comprehensive exploration of various models, which could offer a more nuanced understanding of their strengths and weaknesses in diabetes prediction.

In conclusion, the literature reviewed in this paper shows that ML techniques can predict diabetes with high accuracy. Logistic regression, Decision Trees, SVM, Random Forest, and ANN are the most commonly used ML techniques for diabetic prediction. Furthermore, using a combination of different dataset features as predictors can improve the performance of these techniques. However, more research is needed to investigate the impact of different dataset features on the performance of ML techniques for diabetic prediction.

Our study takes a comprehensive and methodologically rigorous approach to explore the effectiveness of various ML algorithms for predicting and managing diabetes. Unlike some previous works that focused on specific algorithms or lacked detailed feature analysis, we extend our investigation across 11 diverse ML models and employ four distinct real-world datasets, each with unique characteristics. By systematically evaluating predictive performance and feature importance across different datasets, our study aims to enhance the comparability and generalization of results, mitigating the issue of inconsistent dataset characteristics encountered in previous works. Furthermore, our research addresses the variability in evaluation metrics by examining the impact of performance measures such as accuracy, precision, and recall on model assessment. Our rigorous methodology, involving thorough feature scrutiny and validation across diverse datasets, is designed to strengthen the reliability and robustness of our findings, setting a precedent for more comprehensive and conclusive studies in the field.

3. Methodology

We followed a methodology proposed in several studies [24-27]. This methodology is formed by six phases (1) data loading, (2) data pre-processing, (3) feature selection, (4) model execution, (5) model evaluation, and (6) performance analysis. Figure 1 shows the sequential order of these phases in a schematic manner.

The six phases in our methodology can be described as follows:

(1) Collect data from four distinct datasets related to diabetes prediction.

(2) Preprocess the data by handling missing values and resampling to mitigate class imbalance.

(3) Perform feature selection to identify and retain the most informative attributes.

(4) Train 11 different ML models to predict diabetes using the refined dataset.

(5) Use cross-validation to evaluate each model by quantifying performance metrics, such as accuracy, precision,

recall, f1-Score, and ROC_AUC.

(6) Conduct a comparative analysis to find the best classifier given the most important attributes previously selected.

For our experiments, we use scikit-learn [https://scikit-learn.org/], a ML library for Python. Our experiments were executed inside Google Colab [https://colab.research.google.com], a platform for data science and machine learning. The next section provides a detailed description of the processes involved in each of the six phases.

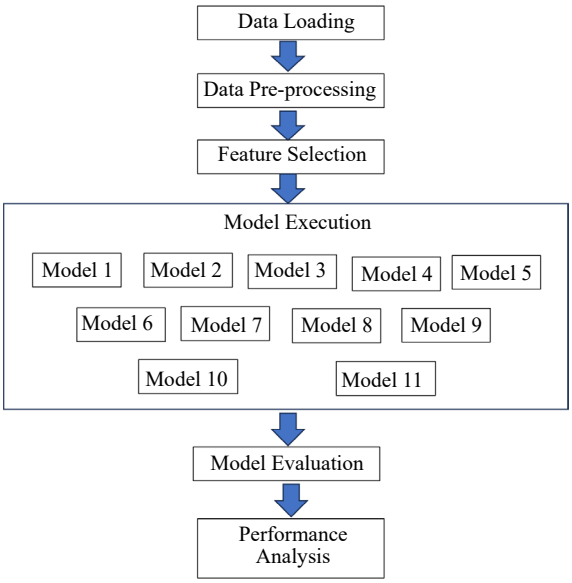


Figure 1. Methodology for this study

4. Performance evaluation

4.1 *Details of datasets*

For our research, we identified four datasets used for diabetes prediction/classification. Each of these datasets contains instances with different attributes of patients and an attribute for the classes of interest. Table 1 shows a summary of these datasets.

Table 1. Summary of diabetes datasets

Dataset	# of Attributes	# of Instances	Classes
Mendeley	12	1,000	N, P, Y
PID	8	768	0, 1
Diabetes Early Stage (DES)	16	520	Negative, Positive
Vanderbilt	15	390	No diabetes, Diabetes

4.1.1 *Mendeley dataset*

This is a publicly accessible dataset [https://data.mendeley.com/datasets/wj9rwkp9c2/1] published in July 2020 by the University of Information Technology [28]. To construct this dataset, the researchers utilized data from Iraqi patients receiving care at the Medical City Hospital laboratory and the Specialized Center for Endocrinology and Diabetes at Al-Kindy Teaching Hospital. The dataset on diabetes was created by systematically reviewing patient files and extracting

relevant information, such as medical history, laboratory analysis results, and patient characteristics. The dataset includes 1,000 instances with 12 attributes: No. of Patient, Age, Gender, Creatinine ratio (Cr), Body Mass Index (BMI), Urea, Cholesterol (Chol), Fasting lipid profile (LDL, VLDL), Triglycerides (TG), HDL Cholesterol, and HBA1C. Similarly, it contains the classes for diabetic (Y), non-diabetic (N), and pre-diabetic(P). For this study, we consider the pre-diabetic (P) class as part of the diabetic (Y) class. For the positive class, we have 897 instances and 103 for the negative. Figure 2(a) shows the distribution of the final classes.

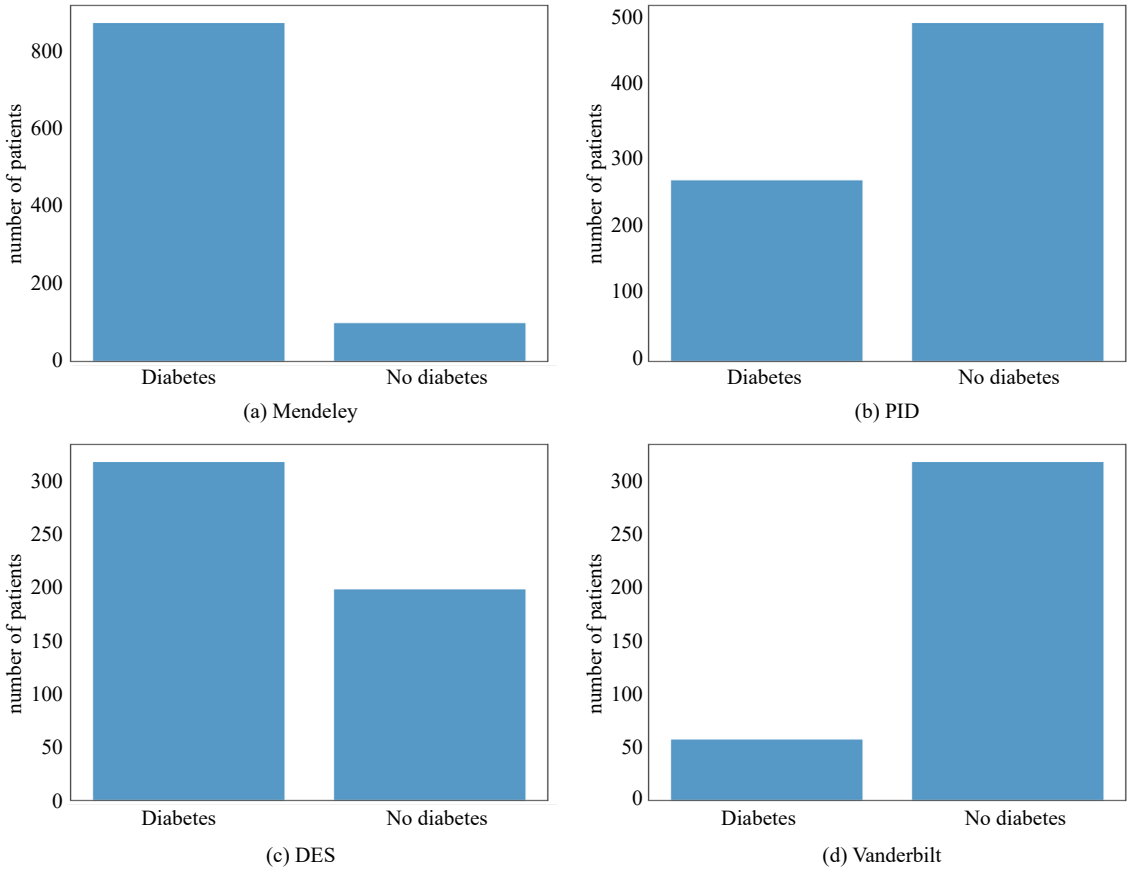


Figure 2. Class distribution for the different datasets

4.1.2 *Pima Indians Diabetes (PID) dataset*

This public dataset [https://data.world/data-society/pima-indians-diabetes-database], originally from the National Institute of Diabetes and Digestive and Kidney Diseases [29], is designed to predict whether a patient has diabetes based on diagnostic measurements. The dataset contains information on 768 patients (i.e., instances). It includes eight characteristics and diagnostic measurements, including pregnancies, plasma glucose concentration, blood pressure, skin thickness, insulin levels, body mass index (BMI), diabetes pedigree function, and age.

It is important to note that the dataset is specifically selected to include only female patients of Pima Indian heritage who are at least 21 years old. This specific population was chosen due to the high incidence of diabetes in this group. The dataset also includes a class variable, which indicates whether the patient has diabetes or not (i.e., 0 or 1). There are 268 instances for the positive class and 500 for the negative. Figure 2(b) shows the distribution of the final classes.

4.1.3 *Diabetes Early Stage (DES) dataset*

This public dataset [https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset]

comprises reports of diabetes-related symptoms from 520 individuals. It includes data on symptoms that may indicate the presence of diabetes and demographic information on the individuals surveyed. The dataset was created by conducting a direct questionnaire with individuals who have recently been diagnosed with diabetes or who are non-diabetic but present with one or more diabetes-related symptoms. The data was collected from patients at the Sylhet Diabetes Hospital in Bangladesh [30].

The dataset contains 16 attributes: Age, Sex, Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, Itching, Irritability, delayed healing, partial paresis, muscle stiffness, Alopecia, Obesity. All of these attributes have categorical values, with "Yes" indicating the presence of a symptom and "No" indicating the absence of a symptom. The dataset also includes two class variables used to determine whether the patient is at risk of developing diabetes (positive) or not (negative). There are 320 instances for the positive class and 200 for the negative. Figure 2(c) shows the distribution of the final classes.

4.1.4 *Vanderbilt dataset*

This public dataset [https://data.world/informatics-edu/diabetes-prediction] is based on a study of rural African Americans in Virginia [22]. There are 390 data samples with both male and female patients. It consists of 15 features that help predict diabetes, including Cholesterol, Glucose, HDL Chol, Chol/HDL ratio, Age, Gender, Height, Weight, BMI, Systolic BP, Diastolic BP, waist, hip, and Waist/hip ratio. Except for Gender, which is categorical (i.e., male and female), the other attributes are numerical. The dataset includes two class variables, "Diabetes" and "No diabetes". There are 60 instances for the positive class and 330 for the negative. Figure 2(d) shows the distribution of the final classes.

4.2 *Experimental setup*

The datasets used in this study contain a class feature. This feature contains binary values that indicate if a patient (i.e., instance) has diabetes or not. Therefore, in our study, we are interested in different ML algorithms for classification.

For this study, we selected 11 of the most commonly used algorithms, which were grouped into six categories. These categories are not mutually exclusive; some algorithms can belong to multiple categories. Table 2 shows the categories and the classifiers used in this study.

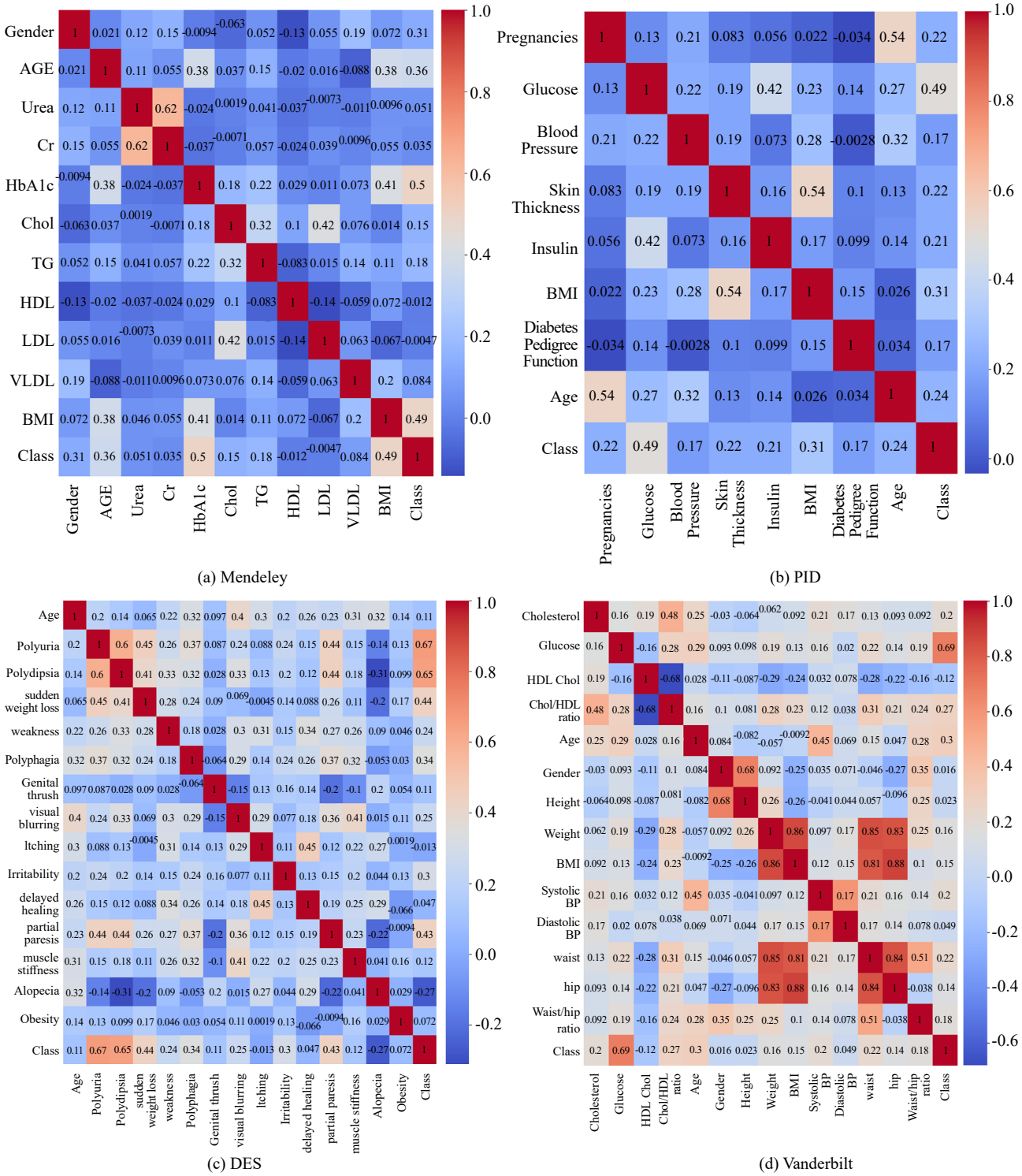
Table 2. Classifiers used in this study

Category	Algorithm
Linear Classifiers	Logistic Regression [31] Quadratic Discriminant Analysis [32] SVC with linear kernel [33]
Non-linear Classifiers	Kernel SVM [34] Multi-Layer Perceptron [35]
Probabilistic Classifiers	Naive Bayes Classifier [36]
Instance-based Classifiers	K-Nearest Neighbors [37]
Tree-based Classifiers	Decision Tree Classifier [38] Random Forest Classifier [39]
Ensemble-based Classifiers	XGBoost [40] AdaBoost [41]

We removed the patient ID or patient number columns for the datasets Mendeley, DES, and Vanderbilt. Similarly, for these datasets, we changed the values for the classes to binary values 0 and 1 (i.e., 0 = no diabetes, 1 = diabetes). The PID dataset had missing values in different columns. To address this problem, we used a Mean imputation strategy. In other words, we replaced each missing value with the mean of the observed values for that column.

Feature engineering is crucial for enhancing the data used to train machine learning models. Removing highly correlated features is a key aspect of feature engineering. Diabetes datasets often contain multiple measurements that

exhibit high correlation, such as fasting plasma glucose and HbA1c [42]. Retaining these redundant features can distort and dilute the importance scores during model training. Furthermore, eliminating these highly correlated features mitigates the risk of model overfitting. Therefore, another data preprocessing step is analyzing the correlation between variables.



(a) Mendeley (b) PID

(c) DES (d) Vanderbilt

Figure 3. Correlation heatmaps for the different datasets

Figure 3 shows the correlation heat maps for the four datasets. A correlation heat map displays the correlation coefficients between multiple variables or features in a dataset. In this figure, red indicates a strong positive correlation, and purple indicates a strong negative correlation. The heat map makes it simple to determine which variables are strongly correlated with one another and which are not. Furthermore, the heat map can be very useful in detecting multicollinearity [43], a phenomenon in which two variables coexist. Figure 3(d) shows that in the Vanderbilt dataset there are columns with high correlation (i.e., multicollinearity), with 85% or above, that can lead to unstable and unreliable estimates. The columns "Waist", "Hip", and "Weight" are highly correlated because they are all measures of body size and shape. Research [44] has found that the waist-to-hip ratio (WHR) effectively predicts if a person is at risk of death from heart disease, cancer, diabetes, or any other cause. Therefore, given that the dataset already has a column "Waist/hip ratio", we decided to drop the columns "Waist", "Hip", and "Weight" from this dataset.

We used SMOTE [45] to address the class imbalance in the datasets. SMOTE is commonly used in machine learning, particularly in classification tasks, to improve the performance of models on imbalanced datasets. We used a value of 5 to define the neighborhood of samples and to resample all classes but the majority class.

We used a grid search [46] approach to find each model's best hyperparameter settings. Grid search allows us to explore a range of possible hyperparameter settings systematically and then select the model that yields the best performance. This process can reduce the time and effort needed to optimize a model's parameters and improve the predictions' accuracy.

In summary, these techniques together help prevent overfitting by ensuring models are trained on representative, unbiased, and diverse data with relevant, non-redundant features. Cross-validation helps assess generalization performance, while techniques like SMOTE address class imbalance issues. Additionally, hyperparameter tuning through grid-search optimizes models for unseen data. Together, these methods will support our results and the ability to apply the findings more widely.

4.3 *Performance metrics*

We used k-fold cross-validation [47] in our experiments to evaluate the performance of each ML model. The benefits of using k-fold cross-validation lie in its ability to provide a robust and comprehensive assessment of model performance, mainly when working with limited datasets, by iteratively partitioning the data into training and validation sets, thereby reducing the variance in performance estimation [48]. In our experiments, we adopted a value of k = 10, a commonly utilized parameter in similar studies [49]. Figure 4 shows the k-fold configuration strategy followed in this study. With this configuration, we systematically split the data into ten parts: nine for training the model and one for testing. The process is repeated ten times, with each part used for testing once.

Figure 4. k-fold cross validation with k = 10

Different performing metrics were used after performing this cross-validation. Table 3 shows a description of the metrics we used in this study. Each of these metrics provides a different perspective on the model's performance and it is essential to consider all of them when evaluating a ML model. In some cases, good accuracy might not be enough to consider a model good, and precision, recall, and other metrics might be more important, especially with highly imbalanced data.

Table 3. Metrics used for evaluating the models

Metric	Description
Accuracy	It is a measure of the proportion of correct predictions made by the model, calculated as the number of correct predictions divided by the total number of predictions.
Precision	It is a measure of the proportion of positive predictions that are actually correct, calculated as the number of true positive predictions divided by the sum of the true positive and false positive predictions.
Recall	It is a measure of the proportion of actual positive cases that are correctly identified by the model, calculated as the number of true positive predictions divided by the sum of the true positive and false negative predictions.
f1-Score	It is a measure of the balance between precision and recall, calculated as the harmonic mean of Precision and Recall, with a higher score indicating a better balance between the two.
ROC_AUC	It is a measure of the model's ability to distinguish between positive and negative classes, with a higher score indicating a better model performance. It is calculated by plotting the true positive rate (recall) against the false positive rate at different classification thresholds and measuring the area under the curve.

4.4 *Results and discussion*

In this section, we present the findings of our study, including the performance of the different ML algorithms on the diabetes datasets. Table 4 shows the best parameters obtained by using grid search for each of the models that had the highest accuracy for the different datasets when using the top 2, top 3, and all features (refer to Sections 4.4.2, 4.4.3, and 4.4.4).

Table 4. Best parameters

Dataset	Top 2 features	Top 3 features	All features
Mendeley	criterion: gini max_depth: 6 max_features: auto n_estimators: 100	gamma: 0 learning_rate: 0.01 max_depth: 3 subsample: 1	criterion: entropy max_depth: 5 max_features: auto n_estimators: 100
PID	C: 100 gamma: 1 kernel: rbf	C: 10 gamma: 1 kernel: rbf	C: 100 gamma: 1 kernel: rbf
DES	C: 0.001 max_iter: 100 penalty: l2 solver: newton-cg	C: 1 gamma: 1 kernel: rbf	criterion: gini max_depth: 8 max_features: sqrt n_estimators: 600
Vanderbilt	C: 1,000 gamma: 0.1 kernel: rbf	activation: relu alpha: 0.0001 hidden_layer_sizes: (50, 100, 50) learning_rate: invscaling solver: adam	activation: relu alpha: 0.0001 hidden_layer_sizes: (50, 50, 50) learning_rate: constant solver: adam

4.4.1 *Most important attributes*

We used the Random Forest (RF) [50] algorithm to get the most important attributes. We decided to use this technique, given that existing literature [51-53] provides substantial evidence through comparative assessments that RF represents an effective data-driven approach for identifying relevant input features across various use cases.

RF is a powerful method that can identify the most informative features in a dataset. The feature importance measure in RF is calculated based on the decrease of the impurity in the data resulting from using the feature to split the data. RF creates multiple decision trees, each trained on a different subset of the data, and then averages the results from all the trees. The decrease in impurity is calculated as the weighted average of the decrease over all the decision trees in the forest. This measure allows RF to identify the most informative features by separating the data into different classes. The feature with the highest decrease in impurity is considered the most important feature.

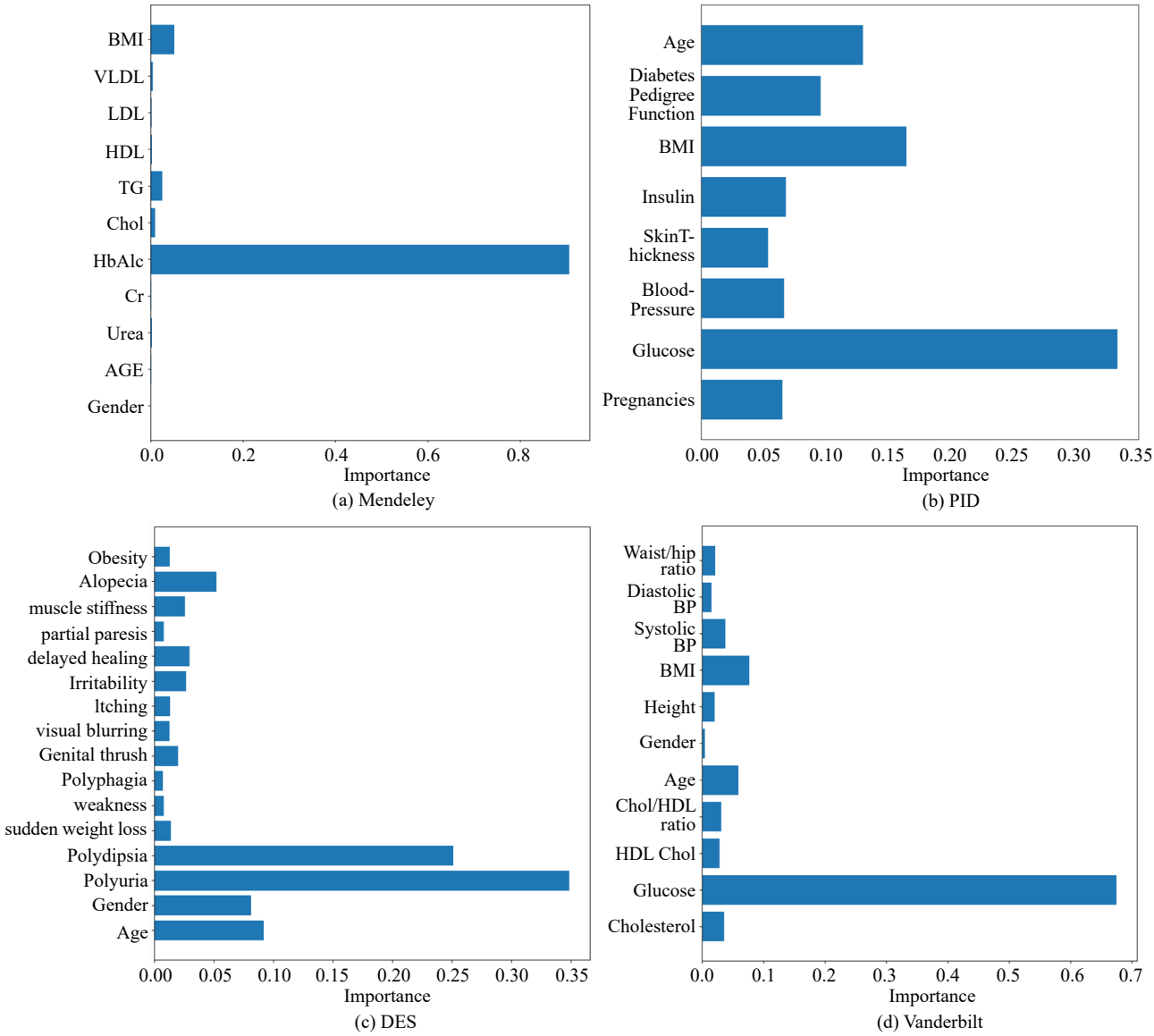


Figure 5. Most important features for each dataset

For this study, we focus on finding the top 2 and top 3 most important features that could determine if a patient

has diabetes or not. Figure 5 shows the most important features of each dataset. Upon a comprehensive review of the generated plots for each dataset, a notable observation was made regarding the Mendeley dataset. Specifically, three features emerged as particularly representative. Upon reviewing the plots generated for each dataset, we observed that three features are particularly representative in the Mendeley dataset. To maintain consistency across all datasets, we chose three as the maximum number of features for comparison. Other works related to disease prediction have presented a similar process [54-55].

For Mendeley (Figure 5(a)), the top 3 features are "HbA1c", "BMI", and "TG", in that order of importance. HbA1c, also known as glycated hemoglobin, is a blood test used to measure the average blood sugar levels over the past 2-3 months in people with diabetes. A high HbA1c level indicates poor blood sugar control and an increased risk of diabetes-related complications [56]. Body Mass Index (BMI) is often used as an indicator of the risk of developing diabetes or as a way to monitor diabetes management [57]. Finally, high levels of triglycerides (TG) in the blood are associated with an increased risk of developing type 2 diabetes and heart disease [58].

For PID (Figure 5(b)), the top 3 features in order of importance are "Glucose", "BMI", and "Age". The "Glucose" feature measures the amount of sugar in the blood after an OGTT test [59]. After this test, high levels of sugar in the blood may show that a person has diabetes or is at risk of developing it. The "BMI" feature, was described above. Lastly, the "Age" feature is relevant in this dataset, mainly because as people get older, their body's ability to use insulin decreases, which can lead to diabetes.

For DES (Figure 5(c)), the top 3 most important features in order of importance are "Polyuria", "Polydipsia", and "Age". "Polyuria" and "Polydipsia" are symptoms of diabetes that are related to high blood sugar levels [60]. In this dataset, the feature "Age" is also a factor for diabetes. We have mentioned the importance of this feature previously.

Finally, for Vanderbilt, we have "Glucose", "BMI", and "Age" as the top 3 features (Figure 5(d)) in that order of importance. "Glucose" represents the values obtained for a Fasting Blood Sugar (FBS) test [61]. This test measures the amount of glucose in a person's blood after fasting for at least 8 hours. The features "BMI" and "Age" were described previously.

After getting the most important features for the four datasets, we can see that "BMI" appears as an important factor in three of them (i.e., DES does not contain a "BMI" attribute). A similar situation happens with "Age". Interestingly, "Age" is not an important factor in Mendeley. This finding highlights the importance of these factors (i.e., "BMI" and "Age") in developing and managing diabetes. Similarly, these results emphasize the need for people to maintain a healthy BMI and to be aware of the risk of diabetes as they age.

4.4.2 *Performance evaluation using the top 2 attributes*

We evaluated the effectiveness of eleven machine learning algorithms on the top 2 attributes obtained in Section 4.4.1. We used a k-fold cross-validation method for the different datasets with k set to 10.

For the Mendeley dataset, the top 2 attributes are HbA1c and BMI. Table 5 shows that the Random Forest Classifier performed the best in terms of accuracy, recall, f1-Score, and ROC_AUC, with 98.83%, 98.04%, 98.84%, and 98.85%, respectively. The XGBoost model also performed well with 98.77% accuracy, 100% precision, 97.59% recall, 98.77% f1-Score, and 98.79%. Its recall was only lower than the Random Forest and K-Nearest Neighbor models. The Kernel SVM, Ada Boost, K-Nearest Neighbor, and Multi-Layer Perceptron also performed well with accuracy scores above 98%. The Naive Bayes and Quadratic Discriminant Analysis models had lower accuracy scores but still performed reasonably well. The average accuracy considering all models is 97.59%, with a median of 98.27%. To sum up, the Random Forest Classifier and XGBoost models are the best performers among the models tested.

Glucose and BMI are the top 2 attributes used in the PID dataset. Table 6 shows that the Kernel SVM and the Random Forest Classifier models performed best. Kernel SVM presented the highest accuracy, recall, f1-Score, and ROC_AUC.

Only Naive Bayes presented the best precision. Random Forest Classifier shows the second-best percentages after the Kernel SVM model. Overall, the accuracy obtained from this dataset using only two features is low. The average accuracy considering all models is 73.65% with a median of 73.6%.

Table 5. Performance analysis for the classifiers on top 2 attributes for the Mendeley dataset

Predictive Model	% accuracy	% precision	% recall	% f1-Score	% ROC_AUC
Random Forest Classifier	98.83	99.67	98.04	98.84	98.85
Logistic Regression	97.16	99.33	95.01	97.10	97.16
XGBoost	98.77	100.00	97.59	98.77	98.79
Decision Tree Classifier	97.77	99.17	96.38	97.73	97.82
SVC (linear kernel)	97.49	99.66	95.32	97.43	97.49
Kernel SVM	98.27	99.67	96.88	98.25	98.26
Naive Bayes	94.87	99.40	90.25	94.58	94.83
AdaBoost Classifier	98.61	99.67	97.59	98.61	98.63
K-Nearest Neighbor	98.38	99.00	97.8	98.39	98.39
Quadratic Discriminant Analysis	94.76	100.00	89.48	94.42	94.74
Multi-Layer Perceptron	98.61	99.64	97.59	98.60	98.63

Table 6. Performance analysis for the classifiers on top 2 attributes for the PID dataset

Predictive Model	% accuracy	% precision	% recall	% f1-Score	% ROC_AUC
Random Forest Classifier	75.60	74.03	78.86	76.21	75.58
Logistic Regression	71.90	73.33	68.90	70.76	71.87
XGBoost	73.80	73.46	74.79	73.93	73.87
Decision Tree Classifier	72.30	69.31	79.87	73.86	72.23
SVC (linear kernel)	72.60	74.91	68.56	71.31	72.72
Kernel SVM	76.20	74.04	80.99	77.20	76.22
Naïve Bayes	73.60	76.35	68.72	72.12	73.69
AdaBoost Classifier	73.00	70.71	78.15	74.03	72.89
K-Nearest Neighbor	73.70	71.55	79.18	74.92	73.81
Quadratic Discriminant Analysis	73.00	75.70	68.13	71.47	72.97
Multi-Layer Perceptron	74.50	72.52	79.24	75.55	74.66

Table 7 presents the performance results for the algorithms in the DES dataset. The top 2 attributes used are “Polyuria” and “Polydipsia”. The results show that most of the algorithms have relatively high accuracy, pprecision, Eecall, and f1-Score scores, with values ranging from 87% to 90% and the ROC_AUC scores ranging from 87% to 88%. The Logistic Regression model presents the overall performance in accuracy, precision, f1-Score, and ROC_AUC. Only the K-Nearest Neighbor model shows better recall. Random Forest Classifier has the second-best overall performance. On the other hand, the K-Nearest Neighbor algorithm has low accuracy, precision, recall, and F1-Score score. The accuracy results obtained from this dataset using only two features are better than those obtained from the PID dataset but not as good as those from the Mendeley dataset. The average accuracy of all models was 84.67%, with a median of 88.12%.

Table 7. Performance analysis for the classifiers on top 2 attributes for the DES dataset

Predictive Model	% accuracy	% precision	% recall	% f1-Score	% ROC_AUC
Random Forest Classifier	88.59	90.98	85.93	88.07	88.74
Logistic Regression	88.75	91.19	85.93	88.19	88.89
XGBoost	88.59	90.71	85.93	88.04	88.77
Decision Tree Classifier	87.50	88.87	85.93	87.12	87.56
SVC (linear kernel)	87.97	89.75	85.93	87.51	88.08
Kernel SVM	88.12	90.01	85.93	87.63	88.26
Naïve Bayes	88.44	90.55	85.93	87.90	88.62
AdaBoost Classifier	88.28	90.20	85.93	87.79	88.36
K-Nearest Neighbor	50.00	50.00	100.00	66.55	50.00
Quadratic Discriminant Analysis	87.03	87.80	85.93	86.66	87.16
Multi-Layer Perceptron	88.12	89.88	85.93	87.62	88.29

Finally, we used “Glucose” and “BMI” as the top 2 features for the Vanderbilt dataset. Table 8 shows that most algorithms performed well, with accuracy scores ranging from 81.36% to 90.45%. The kernel SVM model obtained the highest value for accuracy, recall, f1-Score, and ROC_AUC. The Naive Bayes model had the best precision with 94.03%, although it presented the worst recall with 66.82%. The models using this dataset obtained an average accuracy of 87.56% with a median of 88.79%. These results are better than those obtained for the PID and DES datasets.

Table 8. Performance analysis for the classifiers on top 2 attributes for the Vanderbilt dataset

Predictive Model	% accuracy	% precision	% recall	% f1-Score	% ROC_AUC
Random Forests Classifier	89.70	88.78	91.02	89.77	89.45
Logistic Regression	86.67	90.91	81.00	85.56	86.54
XGBoost	89.09	88.60	89.94	89.14	88.94
Decision Tree Classifier	87.58	87.95	87.20	87.47	87.10
SVC (linear kernel)	86.36	92.07	79.55	85.28	86.42
Kernel SVM	90.45	88.5	93.30	90.73	90.50
Naïve Bayes	81.36	94.03	66.82	77.92	81.32
AdaBoost Classifier	89.70	87.62	92.41	89.87	89.32
K-Nearest Neighbor	88.79	87.2	90.36	88.70	88.57
Quadratic Discriminant Analysis	83.94	92.65	74.29	82.20	84.13
Multi-Layer Perceptron	89.55	89.37	90.11	89.61	89.44

The results showed that using the top two attributes produced a high accuracy of the models in the Mendeley dataset, with an average accuracy of 97.59%. However, the accuracy was lower for the PID dataset, with a score lower than 80%. Comparably, the average accuracy achieved in both the DES and Vanderbilt datasets, while surpassing the accuracy of the PID dataset, did not attain the levels observed in the Mendeley dataset. Out of the 11 models tested, the Kernel SVM performed the best in terms of accuracy, recall, f1-Score, and ROC_AUC in the PID and Vanderbilt datasets. Although the Random Forest Classifier was the top performer in the Mendeley dataset, the overall performance of the Kernel SVM was still comparable in the same dataset.

4.4.3 *Performance evaluation using the top 3 attributes*

As in the previous section, we used a k-fold cross-validation strategy to evaluate the performance of the eleven classifiers using the top 3 attributes obtained in Section 4.4.1.

Table 9 shows the results for the Mendeley dataset. The top 3 attributes for this dataset are "HbA1c", "BMI", and "TG". The highest accuracy was achieved by XGBoost (98.89%), followed by Random Forest Classifier (98.83%) and AdaBoost Classifier (98.83%). The lowest accuracy was obtained by Naive Bayes (94.20%). Regarding precision, Kernel SVM performed the best (99.89%), followed by Random Forest Classifier (99.69%). The lowest precision was observed for Naive Bayes (97.18%). For recall, the AdaBoost Classifier performed the best (98.58%), whereas the lowest recall was observed for Quadratic Discriminant Analysis (90.39%). XGBoost (98.89%) performs the best for the f1-Score. The lowest f1-Score was observed for Naive Bayes (94%). Regarding ROC_AUC, XGBoost performed the best (98.91%), and Naive Bayes (94.16%) had the lowest score. Overall, the results suggest that XGBoost is the best-performing model based on the metrics used for this dataset.

Table 9. Performance analysis for the classifiers on top 3 attributes for the Mendeley dataset

Predictive Model	% accuracy	% precision	% recall	% f1-Score	% ROC_AUC
Random Forest Classifier	98.83	99.69	98.03	98.85	98.84
Logistic Regression	97.10	98.65	95.55	97.05	97.10
XGBoost	98.89	99.78	98.04	98.89	98.91
Decision Tree Classifier	98.22	99.66	96.81	98.21	98.23
SVC (linear kernel)	97.49	99.05	95.89	97.43	97.51
Kernel SVM	98.44	99.89	96.97	98.40	98.43
Naïve Bayes	94.20	97.18	91.05	94.00	94.16
AdaBoost Classifier	98.83	99.14	98.58	98.85	98.83
K-Nearest Neighbor	98.33	99.54	97.11	98.31	98.33
Quadratic Discriminant Analysis	94.31	98.06	90.39	94.05	94.31
Multi-Layer Perceptron	98.77	99.01	98.57	98.78	98.78

In Table 10, we can find the results for the top 3 attributes in the PID dataset. These attributes are "Glucose", "BMI", and "Age". The results show that the best-performing algorithm in terms of accuracy is the Kernel SVM model, with a score of 79%. The Quadratic Discriminant Analysis had a lower accuracy score of 71.7%. Regarding precision, the Naive Bayes algorithm presented the highest score with 77.08%. The Decision Tree Classifier had the lowest precision score, 72.37%. For recall, the K-Nearest Neighbor also had the highest score with 86.16%. The Quadratic Discriminant Analysis had a lower recall score of 66.67%. The K-Nearest Neighbor had the highest f1-Score of 80.1%. The Quadratic Discriminant Analysis had the lowest f1-Score of 69.79%. Finally, the Kernel SVM model had the highest ROC_AUC score. The Quadratic Discriminant Analysis had the lowest ROC_AUC score of 71.89%.

The top 3 features for the DES dataset are "Polyuria", "Polydipsia", and "Age". Table 11 shows the results for the algorithms when using these features. The table shows that the Kernel SVM model achieves the highest accuracy, 89.69%. K-Nearest Neighbor achieves the lowest accuracy, with an accuracy of 56%. Regarding precision, the highest is achieved by the Kernel SVM with 98.47%, and the K-Nearest Neighbor achieves the lowest with 79.50%. Three models present the highest recall. Logistic Regression, SVC with a linear kernel, and Quadratic Discriminant Analysis, all with 85.93%. The Kernel SVM model achieves the lowest recall with 80.63%. The Random Forest Classifier achieves the highest f1-Score with 88.63%. K-Nearest Neighbor achieves the lowest f1-Score with 82.19%. Finally, the highest ROC_AUC is achieved by the Kernel SVM model with a ROC_AUC of 89.73%. K-Nearest Neighbor achieves the lowest ROC_AUC with 81.52%. In general, it can be seen that the Kernel SVM model performs the best in terms of accuracy, precision, and ROC_AUC.

Table 10. Performance analysis for the classifiers on top 3 attributes for the PID dataset

Predictive Model	% accuracy	% precision	% recall	% f1-Score	% ROC_AUC
Random Forest Classifier	78.10	75.45	83.46	79.11	78.16
Logistic Regression	72.70	74.19	70.00	71.71	72.84
XGBoost	77.20	75.73	79.80	77.61	77.09
Decision Tree Classifier	74.20	72.37	79.31	75.31	74.32
SVC (linear kernel)	72.70	73.96	70.75	71.94	72.94
Kernel SVM	79.00	76.97	82.79	79.66	79.04
Naïve Bayes	74.90	77.08	71.09	73.67	74.96
AdaBoost Classifier	76.30	74.91	79.82	77.03	76.43
K-Nearest Neighbor	78.80	75.21	86.16	80.10	78.96
Quadratic Discriminant Analysis	71.70	74.35	66.67	69.79	71.89
Multi-Layer Perceptron	77.90	76.09	81.41	78.49	77.83

Table 11. Performance analysis for the classifiers on top 3 attributes for the DES dataset

Predictive Model	% accuracy	% precision	% recall	% f1-Score	% ROC_AUC
Random Forest Classifier	89.53	95.32	83.12	88.63	89.59
Logistic Regression	88.28	90.15	85.93	87.76	88.43
XGBoost	89.22	94.34	83.80	88.52	89.22
Decision Tree Classifier	89.53	95.69	82.89	88.61	89.65
SVC (linear kernel)	87.81	89.34	85.93	87.38	87.87
Kernel SVM	89.69	98.47	80.63	88.46	89.73
Naïve Bayes	87.81	89.32	85.93	87.38	87.86
AdaBoost Classifier	88.44	91.35	85.03	87.76	88.60
K-Nearest Neighbor	81.56	79.50	85.44	82.19	81.52
Quadratic Discriminant Analysis	88.12	89.90	85.93	87.62	88.30
Multi-Layer Perceptron	87.81	90.78	84.34	87.06	88.01

Table 12 shows the results for the algorithms when using the top 3 features (i.e., “Glucose”, “BMI”, and “Age”) for the Vanderbilt dataset.

Table 12. Performance analysis for the classifiers on top 3 attributes for the Vanderbilt dataset

Predictive Model	% accuracy	% precision	% recall	% f1-Score	% ROC_AUC
Random Forest Classifier	91.82	91.81	92.18	91.90	91.69
Logistic Regression	86.06	89.17	81.94	85.23	85.92
XGBoost	90.30	89.16	91.45	90.17	90.00
Decision Tree Classifier	89.09	88.05	90.59	89.13	88.96
SVC (linear kernel)	86.52	89.61	83.60	86.08	86.67
Kernel SVM	91.82	90.90	92.98	91.84	91.68
Naïve Bayes	85.45	94.12	76.04	83.84	85.57
AdaBoost Classifier	90.76	87.95	94.84	91.20	90.54
K-Nearest Neighbor	90.91	87.76	94.94	91.10	90.66
Quadratic Discriminant Analysis	85.91	92.92	77.16	84.13	85.72
Multi-Layer Perceptron	93.64	92.52	95.19	93.76	93.55

The top model in terms of accuracy is Multi-Layer Perceptron, with accuracy scores of 93.64%. The model with the lowest accuracy is Naive Bayes, with 85.45%. In terms of precision, again, Multi-Layer Perceptron had the highest score with 92.52%, whereas K-Nearest Neighbor had the lowest with 87.76%. Multi-Layer Perceptron also presents the highest recall with 95.19%, and Naive Bayes the lowest with 76.04%. A similar situation is presented for f1-Score and ROC_AUC. Multi-Layer Perceptron presents the best results, with 93.76% and 93.55%, respectively. Naive Bayes presents the lowest values with 83.84% and 85.57% for f1-Score and ROC_AUC.

The results showed that using the top three attributes produced a high accuracy of the models in the Mendeley dataset, with an average accuracy of 97.58%. The accuracy was lower for the PID dataset, with an average score of 75.7%. The average accuracy obtained by the DES (88%) and Vanderbilt (89.3%) datasets, while higher than the PID dataset, did not show the high values observed in the Mendeley dataset. In general, Multi-Layer Perceptron obtained the best results in all metrics for this dataset.

4.4.4 *Performance evaluation using all attributes*

Analyzing the top 2 and top 3 attributes in a dataset can be helpful for classification by reducing the dimensionality of the data and making it easier to visualize and interpret. Focusing on this small number of attributes makes it easier to understand the relationships between the features and the target variables and identify patterns and trends. However, comparing the results of the top 2 and top 3 attributes with using all attributes in the dataset for a similar task is essential. This comparison is critical since reducing the number of attributes may result in a loss of information and may not accurately represent the full complexity of the data. Similarly, using a subset of the data can result in suboptimal classification performance, as we could miss important relationships between features and the target variable.

In this section, we present the results obtained from the models when using all the attributes for each of dataset. Table 13 shows the results for the Mendeley dataset. The Random Forest Classifier has the highest accuracy, with a score of 99.67%. The worst algorithm in terms of accuracy is Naive Bayes, with 92.75%. The Random Forest Classifier, K-Nearest Neighbor, and Multi-Layer Perceptron have the best precision, scoring 100%. The worst algorithm in terms of precision is Naive Bayes, with 95.49%. As in the previous metrics, the best algorithm in terms of recall is the Random Forest Classifier, with 99.36%. The worst algorithm in terms of recall is Naive Bayes with, 89.81%. One more time, the Random Forest Classifier has the best f1-Score with a score of 99.67%. Similar to other metrics, the worst algorithm in terms off1-Score is Naive Bayes, with 92.53%. Finally, regarding ROC_AUC the Random Forest Classifier and Logistic Regression are the winners with 99.68%. The worst algorithm in terms of ROC_AUC is Naive Bayes, with 92.68%. Overall, the Random Forest Classifier is the algorithm with the highest numbers for all five metrics.

Table 13. Performance analysis for the classifiers on all attributes for the Mendeley dataset

Predictive Model	% accuracy	% precision	% recall	% f1-Score	% ROC_AUC
Random Forest Classifier	99.67	100.00	99.36	99.67	99.68
Logistic Regression	96.71	97.99	95.43	96.68	96.68
XGBoost	99.28	99.90	98.69	99.28	99.28
Decision Tree Classifier	98.22	99.18	97.26	98.21	98.25
SVC (linear kernel)	97.71	99.77	95.62	97.64	97.70
Kernel SVM	98.94	98.78	99.10	98.94	98.93
Naïve Bayes	92.75	95.49	89.81	92.53	92.68
AdaBoost Classifier	99.55	99.89	99.23	99.56	99.56
K-Nearest Neighbor	98.10	100.00	96.20	98.05	98.10
Quadratic Discriminant Analysis	93.87	97.49	90.04	93.59	93.84
Multi-Layer Perceptron	99.22	100.00	98.43	99.20	99.21

Table 14 shows the results for all the models with the PID dataset. The Kernel SVM model has the highest

accuracy, scoring 82.4%. The worst algorithm in terms of accuracy is Naive Bayes, with 72.6%. Similarly, for precision, Kernel SVM was the best model, scoring 81.87%. The worst algorithm in terms of precision is the Decision Tree Classifier, which has a precision of 72.93%. The best algorithm in terms of recall is the K-Nearest Neighbor model, with 90.31%. The worst algorithm in terms of recall is Naive Bayes, with 68.33%. The Multi-Layer Perceptron has the best f1-Score with a score of 82.5%. Similar to other metrics, the worst algorithm in terms off1-Score is Naive Bayes, with 71.21%. Finally, regarding ROC_AUC the Kernel SVM model is the winner, with 82.62%. The worst algorithm in terms of ROC_AUC is Naive Bayes, with 92.68%. Overall, Kernel SVM is the algorithm with the highest numbers in three of the five metrics.

Table 14. Performance analysis for the classifiers on all attributes for the PID dataset

Predictive Model	% accuracy	% precision	% recall	% f1-Score	% ROC_AUC
Random Forest Classifier	80.00	76.75	85.62	80.84	80.00
Logistic Regression	75.5	77.02	73.41	74.67	75.80
XGBoost	81.80	79.15	86.33	82.46	81.78
Decision Tree Classifier	75.60	72.93	81.55	76.75	75.37
SVC (linear kernel)	75.10	76.62	73.05	74.23	75.40
Kernel SVM	82.40	81.87	83.61	82.44	82.62
Naïve Bayes	72.60	74.65	68.33	71.21	72.68
AdaBoost Classifier	79.20	77.48	82.11	79.65	79.21
K-Nearest Neighbor	79.70	74.46	90.31	81.42	79.81
Quadratic Discriminant Analysis	74.70	77.01	70.56	73.21	74.80
Multi-Layer Perceptron	82.10	79.86	85.47	82.50	82.05

Table 15 displays the results of all the models tested on the DES dataset. The Random Forest Classifier emerged as the top performer, while the Naive Bayes model was the weakest. The accuracy scores were 98.12% and 88.44% for Random Forest Classifier and Naive Bayes, respectively. Similarly, the precision scores were 98.7% and 87.72%. Random Forest Classifier also boasted a higher recall score of 97.53%, compared to Naive Bayes' 89.35%. The f1-Score also showed a similar pattern, with Random Forest Classifier scoring 98.07% and Naive Bayes coming in at 88.42%. Finally, regarding ROC_AUC, both models showed a high difference between their results, with the Random Forest Classifier scoring 98.18% and Naive Bayes scoring 88.52%.

Table 15. Performance analysis for the classifiers on ALL attributes for the DES dataset

Predictive Model	% accuracy	% precision	% recall	% f1-Score	% ROC_AUC
Random Forest Classifier	98.12	98.70	97.53	98.07	98.18
Logistic Regression	92.19	92.79	91.01	91.82	92.16
XGBoost	97.34	98.36	96.24	97.26	97.35
Decision Tree Classifier	92.03	94.39	89.43	91.71	92.13
SVC (linear kernel)	93.28	93.91	92.69	93.14	93.23
Kernel SVM	96.72	96.89	96.49	96.64	96.69
Naïve Bayes	88.44	87.72	89.35	88.42	88.52
AdaBoost Classifier	94.69	95.45	93.64	94.48	94.69
K-Nearest Neighbor	95.31	98.28	92.11	94.99	95.34
Quadratic Discriminant Analysis	95.00	98.67	91.32	94.76	95.06
Multi-Layer Perceptron	94.69	95.32	94.00	94.58	94.68

Finally, Table 16 displays the results of all the models tested on the Vanderbilt dataset. The Multi-Layer Perceptron had the best accuracy performance with 96.52%, while the Naive Bayes model had the lowest with 84.55%. For precision, Kernel SVM score the best at 99.71% and K-Nearest Neighbor presented the lowest score with 86.61%. However, the K-Nearest Neighbor showed the best recall with 98.91%. Naive Bayes had the weakest score for the same metric, with 78.84%. The Multi-Layer Perceptron model had the best performance for f1-Score and ROC_AUC, with 96.6% and 96.49% respectively. Naive Bayes obtined the lowest f1-Score with 83.55%, and the Quadratic Discriminant Analysis model obtained the lowest score for ROC_AUC with 86.55%. Overall, the Multi-Layer Perceptron model presented the best overall performance in this dataset.

Table 16. Performance analysis for the classifiers on all attributes for the Vanderbilt dataset

Predictive Model	% accuracy	% precision	% recall	% f1-Score	% ROC_AUC
Random Forest Classifier	93.79	92.30	96.13	94.09	93.84
Logistic Regression	89.70	92.03	86.91	89.34	89.56
XGBoost	89.39	88.60	90.38	89.37	89.12
Decision Tree Classifier	89.85	90.82	89.11	89.79	89.74
SVC (linear kernel)	86.82	88.36	85.10	86.49	86.88
Kernel SVM	96.36	99.71	92.96	96.13	96.32
Naïve Bayes	84.55	89.78	78.84	83.55	87.77
AdaBoost Classifier	89.85	87.83	93.17	90.30	89.70
K-Nearest Neighbor	91.67	86.61	98.91	92.21	91.75
Quadratic Discriminant Analysis	86.82	89.44	82.96	85.99	86.85
Multi-Layer Perceptron	96.52	94.55	98.81	96.60	96.49

The results showed that using all attributes produced an average accuracy above 90% for the Mendeley, DES, and Vanderbilt datasets. The average accuracy for the PID dataset was only 78.06%. The highest average accuracy was obtained in the Mendeley dataset, with 97.6%. Overall, using all features, the best results were obtained using the Random Forest Classifier.

4.4.5 *Overall comparison of classifiers based on the number of features*

In this section, we want to compare the performance regarding the accuracy metric for the 11 algorithms when using the top 2, top 3, and all features for training and testing. By comparing the accuracy metric for different feature subsets, we can identify which features are most significant for accurate predictions and which algorithms perform best with a smaller or larger number of features. This information can help optimize the performance of the machine learning models, reduce the number of features needed for accurate predictions, and increase the models' efficiency.

Figure 6 compares the accuracy of the machine learning models for the Mendeley dataset. The top 3 attributes in this dataset are "HbA1c", "BMI", and "TG". We can see that 54% of the algorithms (n = 6) present a higher accuracy when using all the features. These algorithms are the Multi-Layer Perceptron, AdaBoost Classifier, Kernel SVM, SVC, XGBoost, and Random Forest Classifier. The average accuracy for these algorithms is 99.06%. The Random Forest Classifier presented the best accuracy with 99.67%. From this Figure, we can also see that 36% (n = 5) of the algorithms had better accuracy when using the top 2 features. These algorithms are Quadratic Discriminant Analysis, K-Nearest Neighbor, Naive Bayes, Decision Tree Classifier, and Logistic Regression. The average accuracy for these algorithms is 96.58%. Interestingly, no algorithm performed better than others when using the top 3 features. These results suggest that using the top 3 features did not provide significant additional information beyond the top 2 features or that the additional features may have introduced noise or increased complexity to the model, leading to decreased performance.

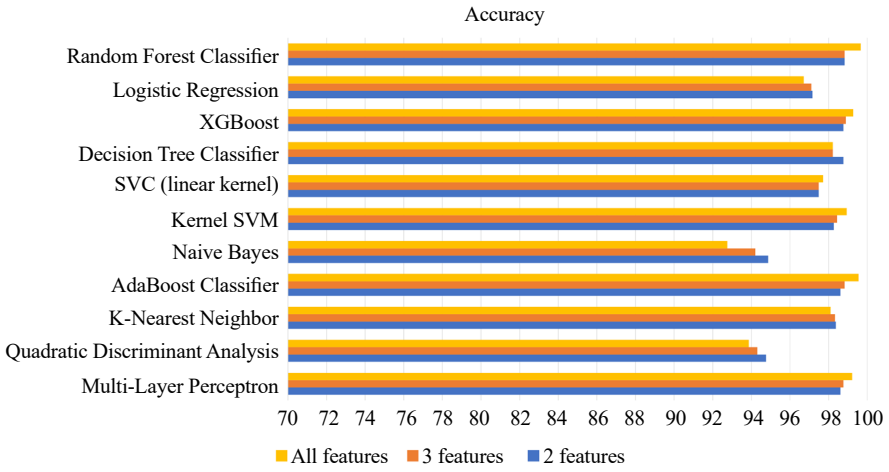


Figure 6. Comparing the accuracy using the top 2, top 3, and all features for the Mendeley dataset

Figure 7 compares the accuracy of the machine learning models for the PID dataset. The top 3 attributes for this dataset are "Glucose", "BMI", and "Age". We can see that the accuracy score for all the classifiers is lower than 84%. Of the different algorithms, 90.9% (n = 10) performed better when using all the features. The average accuracy for these algorithms was 79.61%. Kernel SVM obtained the best accuracy with 82.4%. Only Naive Bayes obtained a better accuracy (74.9%) when using the top 3 features. We see a consistent pattern in the results except for Quadratic Discriminant Analysis and Naive Bayes. The accuracy is worst when using only the top 2 features, slightly better when using the top 3 features, and the best accuracy is achieved when using all the features.

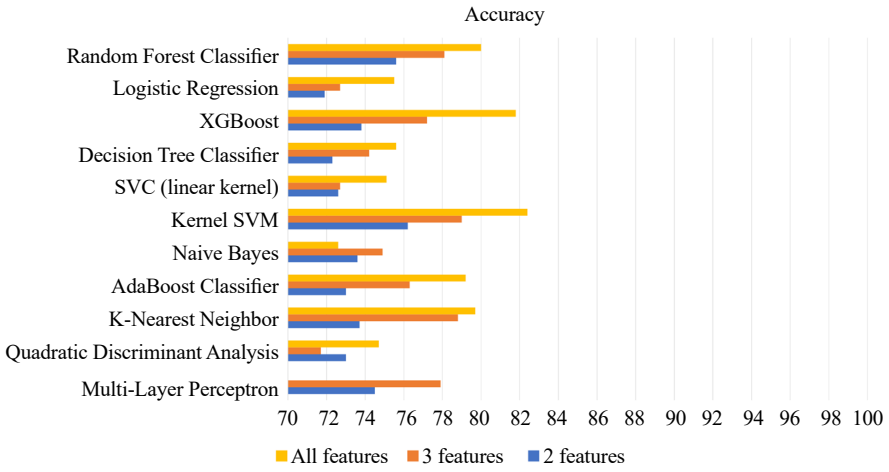


Figure 7. Comparing the accuracy using the top 2, top 3, and all features for the PID dataset

Figure 8 shows the accuracy of the eleven algorithms for the DES dataset. The top 3 attributes for this dataset are "Polyuria", "Polydipsia", and "Age". Except for Naive Bayes, the rest of the algorithms, 90.9% (n = 10), obtained the best accuracy when using all the features. The average accuracy for these algorithms was 94.9%. In the case of Naive Bayes, the algorithm obtained the same accuracy (84.44%) for the top 2 features and all the features. This algorithm also presented the lowest score of the rest of the algorithms for its best performance. As in the previous dataset, we can see a consistent pattern in seven (63.6%) of the algorithms (Quadratic Discriminant Analysis, K-Nearest Neighbor, AdaBoost Classifier, Kernel SVM, Decision Tree Classifier, XGBoost, and Random Forest Classifier). The K-Nearest Neighbor presents the worst accuracy with 50% when using only the top 2 attributes. This result is the lowest score obtained by an

algorithm in all datasets.

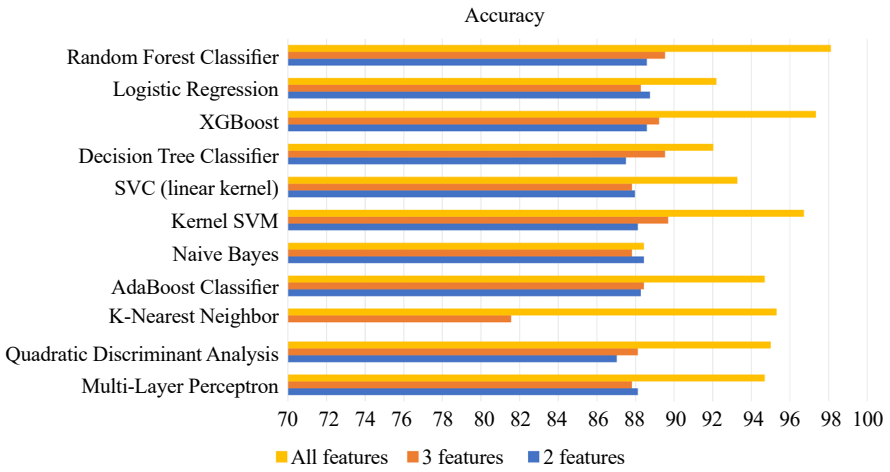


Figure 8. Comparing the accuracy using the top 2, top 3, and all features for the DES dataset

Figure 9 shows the accuracy of the eleven algorithms for the Vanderbilt dataset. The top 3 attributes for this dataset are "Glucose", "MI", and "Age". Most of the algorithms, 72.7% (n = 8), obtained the best accuracy when using all the features. Only XGBoost and AdaBoost Classifier obtained the best results when using three features. None of the algorithms performed the best with just two features. Overall, the Random Forest Classifier presented the best accuracy with 97.79%. Naive Bayes obtained the worst accuracy, 81.36% when using only two attributes.

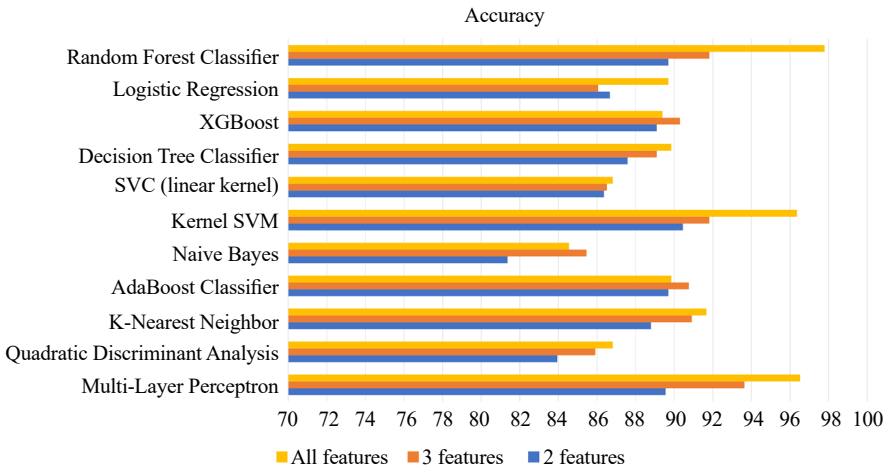


Figure 9. Comparing the accuracy using the top 2, top 3, and all features for the Vanderbilt dataset

4.4.6 *Discussion*

The literature reviewed in this study highlights the efforts of developing effective methods for predicting and preventing diabetes, given its high prevalence and potential for severe complications if not managed properly. Machine learning (ML) techniques have shown promising results in accurately predicting diabetes, with Logistic Regression, Decision Trees, SVM, Random Forest, and ANN being the most commonly used ML techniques. Using a combination of different dataset features as predictors can further improve the performance of these techniques. In this work, we used

four different diabetes datasets, as presented in Section 4.1. Our findings confirm that ML methods hold great potential for predicting and managing diabetes, thereby benefiting public health.

In diabetes prediction, accuracy can vary depending on the specific ML algorithm. Although we focus on predicting diabetes using the different sets of features using accuracy, it is important to note that this metric is just one performance measure.

For the Mendeley dataset, the most important attributes based on the analysis performed using Random Forest were "HbA1c", "BMI", and "TG", in that order. When using only the top 2 attributes, the best accuracy was obtained using the Random Forest Classifier with 98.83%. However, when using the top 3 features, the algorithm presenting the best accuracy was XGBoost, with 98.89%. The Random Forest Classifier was also the algorithm with the best accuracy, with 99.67% when all the attributes were used. Table 17 shows these results and presents the percentage of increment in the accuracy gained using the top 2, 3, and all features. The table shows that the percentage of increment in accuracy obtained from using the top 3 to use all features is not significant. The results show less than a 1% gain in accuracy by using all features.

Table 17. Percentage of increment based on the accuracy for the top 2, 3, and all features for the Mendeley dataset

	2 features	3 features	All features
Algorithm	Random Forest Classifier	XGBoost	Random Forest Classifier
accuracy	98.83	98.89	99.67
% of increment	-	0.061	0.789

In the PID dataset, the most important attributes obtained by the Random Forest technique were "Glucose", "BMI", and "Age", in that order. For this dataset, the algorithm with the best performance, using the top 2, 3, and all features, was Kernel SVM. However, the percentage of accuracy obtained in this dataset was lower than in the others. Table 18 shows these results and presents the percentage of increment in the accuracy gained using the top 2, 3, and all features. We can see a 4.3% accuracy increment from using all the features compared to using only the top 3.

Table 18. Percentage of increment based on the accuracy for the top 2, 3, and all features for the PID dataset

	2 features	3 features	All features
Algorithm	Kernel SVM	Kernel SVM	Kernel SVM
accuracy	76.2	79	82.4
% of increment	-	3.67	4.3

After applying the Random Forest algorithms, the most important attributes for the DES dataset were "Polyuria", "Polydipsia", and "Age", in this order of importance. Three algorithms presented the best performance. Logistic Regression had the best accuracy when using the top 2 attributes, Kernel SVM for the top 3 attributes, and Random Forest Classifier when using all attributes. Table 19 shows that using all features gave a 9.4% accuracy increment compared with only using the top 3 features. This gain is considerable given that the accuracy obtained using only the top 2 or 3 features is less than 90%.

Table 19. Percentage of increment based on the accuracy for the top 2, 3, and all features for the DES dataset

	2 features	3 features	All features
Algorithm	Logistic Regression	Kernel SVM	Random Forest Classifier
accuracy	88.75	89.69	98.12
% of increment	-	1.06	9.4

Finally, for the Vanderbilt dataset, the most important attributes found by the Random Forest algorithm were "Glucose", "BMI", and "Age". Three algorithms presented the best performance for accuracy. Kernel SVM obtained the best results when using the top 2 features. The Multi-Layer Perceptron algorithm was the best when the top 3 features were used. Ultimately, the Random Forest Classifier obtained the best results using all the features. Table 20 compares these results and shows the increment in accuracy obtained using the different features. We can see a 4.43% accuracy increment when using all the features compared to using only the top 3. From these results, we can conclude that using all the features produces an increment in accuracy. However, as the results for the Mendeley dataset show, even with the top 2 attributes, the accuracy for the predictions is high at 98.83%. Therefore, we can infer that feature selection is a crucial step in machine learning, as it reduces the problem's dimensionality and increases the model's interpretability.

Table 20. Percentage of increment based on the accuracy for the top 2, 3, and all features for the Vanderbilt dataset

	2 features	3 features	All features
Algorithm	Kernel SVM	Multi-Layer Perceptron	Random Forest Classifier
accuracy	90.45	93.64	97.79
% of increment	-	3.53	4.43

Our study aimed to evaluate 11 powerful machine learning algorithms on four databases using three sets of features to predict patient outcomes. After thoroughly analyzing 132 results (i.e., 11 × 4 × 3), we found that Kernel SVM scored the highest in five, while Random Forest Classifier achieved the highest accuracy in four. These two algorithms outperformed the others and are highly suitable for accurately predicting patient outcomes.

While this study makes valuable contributions in evaluating machine learning techniques for diabetes prediction, certain limitations remain to be acknowledged. Though comprising real patient data, the datasets have relatively small sample sizes, between 390 to 1,000 instances. Testing on more extensive and more diverse datasets could result in additional insights. Furthermore, our work focused solely on structured tabular data features; incorporating unstructured inputs like clinical notes or medical images using deep learning approaches presents a promising opportunity.

5. Conclusions and future work

Early diagnosis of diabetes can help reduce the mortality rate due to the complications and risks to the patient from this disease. In this paper, we study the performance of 11 different ML algorithms to predict diabetes of individuals using four different datasets: Mendeley dataset, PID dataset, DES dataset, and Vanderbilt dataset. We evaluated and presented the performance analysis of these ML models, selecting the top 2 and 3 attributes obtained from feature selection methods. We used k-fold cross-validation methods to analyze the performance on different performance metrics: accuracy, precision, recall, f1-Score, and ROC_AUC. Our experiments have shown that different algorithms perform differently on various datasets, highlighting the importance of evaluating multiple models to choose the best-performing one. The findings of this study could have significant implications for the medical community in developing accurate predictive models for diagnosing and treating patients. By utilizing the most effective algorithms and

incorporating the right features, we can develop reliable models to improve patient outcomes.

Since a large dataset can provide adequate information for the model to perform better, our suggestions for future work will be to collect and use more data to train and develop more accurate and robust models. We also recommend using different datasets, combining different features of these datasets as predictors to improve the performance of the ML techniques further. Similarly, a study to explore the use of ensembles of different ML algorithms to achieve higher accuracy and using a variety of deep learning methods that have been proved instrumental in other healthcare domains will be the extension of this work. Also, implementing this study in real-life applications, such as wearable devices and web applications, for real-time diabetic prediction and forecasting will be our future interest. Finally, an innovative proposal would use interpretable models that can provide insights into the factors driving diabetes prediction and management.

Acknowledgments

This work was supported partially by the National Science Foundation EPSCoR Program.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. *Frontiers in Genetics*. 2018; 9: 515.

[2] Das SK, Namasudra S, Kumar A, Moparthi NR. AESPNet: Attention enhanced stacked parallel network to improve automatic diabetic foot ulcer identification. *Image and Vision Computing*. 2023; 138: 104809.

[3] Bloomgarden ZT. Diabetes complications. *Diabetes Care*. 2004; 27(6): 1506-1514.

[4] Chaki J, Ganesh ST, Cidham S, Theertan SA. Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. *Journal of King Saud University-Computer and Information Sciences*. 2022; 34(6): 3204-3225.

[5] Alkaragole MLZ, Kurnaz S. Comparison of data mining techniques for predicting diabetes or prediabetes by risk factors. *International Journal of Computer Science and Mobile Computing*. 2019; 8: 61-71.

[6] Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*. 2019; 6(1): 1-19.

[7] Nuankaew P, Chaising S, Temdee P. Average weighted objective distance-based method for type 2 diabetes prediction. *IEEE Access*. 2021; 9: 137015-137028.

[8] Abnoosian K, Farnoosh R, Behzadi MH. Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. *BMC Bioinformatics*. 2023; 24(1): 337.

[9] Dubosson F, Ranvier JE, Bromuri S, Calbimonte JP, Ruiz J, Schumacher M. The open D1NAMO dataset: A multi-modal dataset for research on non-invasive type 1 diabetes management. *Informatics in Medicine Unlocked*. 2018; 13: 92-100.

[10] Hill-Briggs F, Adler NE, Berkowitz SA, Chin MH, Gary-Webb TL, Navas-Acien A, et al. Social determinants of health and diabetes: A scientific review. *Diabetes Care*. 2021; 44(1): 258.

[11] Maliyaem M, Tuan NM, Lockhart D, Muenthong S. A study of using machine learning in predicting COVID-19 cases. *Cloud Computing and Data Science*. 2022; 3(2): 92-99.

[12] Nomura A, Noguchi M, Kometani M, Furukawa K, YonedaT. Artificial intelligence in current diabetes management and prediction. *Current Diabetes Reports*. 2021; 21(12): 61.

[13] Fregoso-Aparicio L, Noguez J, Montesinos L, García-García JA. Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. *Diabetology & Metabolic Syndrome*. 2021; 13(1): 1-22.

[14] Zhuhadar LP, Lytras MD. The application of autoML techniques in diabetes diagnosis: Current approaches,

performance, and future directions. *Sustainability*. 2023; 15(18): 13484.
[15] Juneja A, Juneja S, Kaur S, Kumar V. Predicting diabetes mellitus with machine learning techniques using multi-criteria decision making. *International Journal of Information Retrieval Research*. 2021; 11(2): 38-52.
[16] Chaising S, Temdee P, Prasad R. Weighted objective distance for the classification of elderly people with hypertension. *Knowledge-Based Systems*. 2020; 210: 106441.
[17] Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*. 2022; 18(1): 90-100.
[18] Nguyen BP, Pham HN, Tran H, Nghiem N, Nguyen QH, Do TT, et al. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Computer Methods and Programs in Biomedicine*. 2019; 182: 105055.
[19] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. *Procedia Computer Science*. 2018; 132: 1578-1585.
[20] Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*. 2020; 167: 706-716.
[21] Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*. 2018; 10: 100-107.
[22] Rajendra P, Latifi S. Prediction of diabetes using logistic regression and ensemble techniques. *Computer Methods and Programs in Biomedicine Update*. 2021; 1: 100032.
[23] Gill S, Pathwar P. Prediction of diabetes using various feature selection and machine learning paradigms. In: Gunjan VK, Zurada JM. (eds.) *Modern Approaches in Machine Learning & Cognitive Science: A Walkthrough*. Springer, Cham; 2022. p.133-146.
[24] Amin MS, Chiam YK, Varathan KD. Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*. 2019; 36: 82-93.
[25] Ayon SI, Islam MM, Hossain MR. Coronary artery heart disease prediction: A comparative study of computational intelligence techniques. *IETE Journal of Research*. 2022; 68(4): 2488-2507.
[26] Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*. 2019; 7: 81542-81554.
[27] Fida B, Nazir M, Naveed N, Akram S. Heart disease classification ensemble optimization using genetic algorithm. *2011 IEEE 14th International Multitopic Conference*. Karachi, Pakistan: IEEE; 2011. p.19-24.
[28] Rashid A. Diabetes Dataset. *Mendeley Data*. Available from: doi:10.17632/wj9rwkp9c2.1.
[29] Karegowda AG, Manjunath A, Jayaram M. Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes. *International Journal on Soft Computing*. 2011; 2(2): 15-23.
[30] Islam MMF, Ferdousi R, Rahman S, Bushra HY. Likelihood prediction of diabetes at early stage using data mining techniques. In: Gupta M, Konar D, Bhattacharyya S, Biswas S. (eds.) *Computer Vision and Machine Intelligence in Medical Image Analysis*. Springer: Singapore; 2020. p.113-125.
[31] LaValley MP. Logistic regression. *Circulation*. 2008; 117(18): 2395-2399.
[32] Wu W, Mallet Y, Walczak B, Penninckx W, Massart D, Heuerding S, et al. Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data. *Analytica Chimica Acta*. 1996; 329(3): 257-265.
[33] Hsu CW, Chang CC, Lin CJ. *A Practical Guide to Support Vector Classification*. Taipei, Taiwan; 2003.
[34] Patle A, Chouhan DS. SVM kernel functions for classification. *2013 International Conference on Advances in Technology and Engineering (ICATE)*. Mumbai, India: IEEE; 2013. p.1-9.
[35] Ramchoun H, Ghanou Y, Ettaouil M, Janati Idrissi MA. Multilayer perceptron: Architecture optimization and training. *International Journal of Interactive Multimedia and Artificial Intelligence*. 2016; 4(1): 1-5.
[36] Rish I. An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*. T.J. Watson Research Cente; 2001. p.41-46.
[37] Peterson LE. K-nearest neighbor. *Scholarpedia*. 2009; 4(2): 1883.
[38] Swain PH, Hauska H. The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*. 1977; 15(3): 142-147.
[39] Parmar A, Katariya R, Patel V. A review on random forest: An ensemble classifier. In: Hemanth J, Fernando X, Lafata P, Baig Z. (eds.) *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*. Springer, Cham; 2019. p.758-763.
[40] Liew XY, Hameed N, Clos J. An investigation of XGBoost-based algorithm for breast cancer classification.

Machine Learning with Applications. 2021; 6: 100154.
[41] Bahad P, Saxena P. Study of adaboost and gradient boosting algorithms for predictive analytics. In: Singh Tomar G, Chaudhari NS, Barbosa JLV, Aghwariya MK. (eds.) *International Conference on Intelligent Computing and Smart Communication 2019*. Springer, Singapore; 2020. p.235-244.
[42] Rohlfing CL, Wiedmeyer HM, Little RR, England JD, Tennill A, Goldstein DE. Defining the relationship between plasma glucose and HbA1c: Analysis of glucose profiles and HbA1c in the diabetes control and complications trial. *Diabetes Care*. 2002; 25(2): 275-278.
[43] Alin A. Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*. 2010; 2(3): 370-374.
[44] Borugian MJ, Sheps SB, Kim-Sing C, Olivotto IA, Van Patten C, Dunn BP, et al. Waist-to-hip ratio and breast cancer mortality. *American Journal of Epidemiology*. 2003; 158(10): 963-968.
[45] Fernández A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*. 2018; 61: 863-905.
[46] Jiménez ÁB, Lázaro JL, Dorronsoro JR. Finding optimal model parameters by discrete grid search. In: Corchado E, Corchado JM, Abraham A. (eds.) *Innovations in Hybrid Intelligent Systems*. Springer: Berlin; 2007. p.120-127.
[47] Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S. The 'K'in K-fold cross validation. *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*. Italy; 2012. p.441-446.
[48] Yadav S, Shukla S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. *2016 IEEE 6th International Conference on Advanced Computing (IACC)*. Bhimavaram, India: IEEE; 2016. p.78-83.
[49] Rayyan A, Aburas MG, Al-Mousa A. Uniform resource locator classification using classical machine learning & deep learning techniques. *Cloud Computing and Data Science*. 2023; 4(2): 17-30.
[50] Breiman L. Random forests. *Machine Learning*. 2001; 45(1): 5-32.
[51] Theerthagiri P, Ruby AU. RFFS: Recursive random forest feature selection based ensemble algorithm for chronic kidney disease prediction. *Expert Systems*. 2022; 39(9): e13048.
[52] Wang Y, Xu Y, Yang Z, Liu X, Dai Q. Using recursive feature selection with random forest to improve protein structural class prediction for low-similarity sequences. *Computational and Mathematical Methods in Medicine*. 2021; 2021: 1-9.
[53] Nahar N, Ara F, Neloy MAI, Biswas A, Hossain MS, Andersson K. Feature selection based machine learning to improve prediction of Parkinson disease. In: Mahmud M, Kaiser MS, Vassanelli S, Dai Q, Zhong N. (eds.) *Brain Informatics: BI 2021, Lecture Notes in Computer Science*. Springer, Cham; 2021. p.496-508.
[54] Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. *SN Computer Science*. 2020; 1: 1-6.
[55] Guarneros-Nolasco LR, Cruz-Ramos NA, Alor-Hernández G, Rodríguez-Mazahua L, Sánchez-Cervantes JL. Identifying the main risk factors for cardiovascular diseases prediction using machine learning algorithms. *Mathematics*. 2021; 9(20): 2537.
[56] Weykamp C. HbA1c: A review of analytical and clinical aspects. *Annals of Laboratory Medicine*. 2013; 33(6): 393-400.
[57] Leong KS, Wilding JP. Obesity and diabetes. *Best Practice & Research Clinical Endocrinology & Metabolism*. 1999; 13(2): 221-237.
[58] Wiggin TD, Sullivan KA, Pop-Busui R, Amato A, Sima AA, Feldman EL. Elevated triglycerides correlate with progression of diabetic neuropathy. *Diabetes*. 2009; 58(7): 1634-1640.
[59] Bartoli E, Fra G, Schianca GC. The oral glucose tolerance test (OGTT) revisited. *European Journal of Internal Medicine*. 2011; 22(1): 8-12.
[60] Bellows RT, Van Wagenen WP. The relationship of polydipsia and polyuria in diabetes insipidus. *The Journal of Nervous and Mental Disease*. 1938; 88(4): 417-473.
[61] Ghazanfari Z, Haghdoost AA, Alizadeh SM, Atapour J, Zolala F. A comparison of HbA1c and fasting blood sugar tests in general population. *International Journal of Preventive Medicine*. 2010; 1(3): 187.